基于贝叶斯最小风险的癫痫脑电自动检测算法 *

卫作臣，邹俊忠，张　见，陈兰岚

(华东理工大学，信息科学与工程学院，自动化系，上海 200237)

摘　要：癫痫脑电的自动检测是一个不平衡分类问题。提出一种新的不平衡分类算法，基于增减序列合并周期分割算法提取时域特征，引入随机映射优化了旋转森林的计算效率，进而计算基于海林格距离的贝叶斯最小风险来给出测试样本标签。该算法在 1 s 片段上得到了 90.66%灵敏性，92.52%特异性，F2 分数为 0.9055，并且检出了 98.56%的癫痫发作，检测延迟为 1.32 s，在不平衡的癫痫脑电数据集上表现出了良好的性能，对于癫痫辅助诊断有着极大的临床意义。

关键词：癫痫；时域特征；随机映射；旋转森林；代价敏感；贝叶斯最小风险

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2018.07.0415

Automatic detection of epileptic EEG based on minimum Bayesian risk and rotation forest

Wei Zuochen, Zou Junzhong, Zhang Jian, Chen Lanlan

(*Dept. of Automation*, *School of Information Science & Engineering*, *East China University of Science & Technology*, *Shanghai* 200237, *China*)

Abstract: Automatic detection of epileptic discharges in electroencephalograph (EEG) is an imbalanced classification problem. This paper proposes a novel automatic epileptic EEG detection approach. This study calculates the time domain features based on the merger of increasing and decreasing sequence (MIDS) , and employs Random Projection to improve the complexity of Rotation Forest, as well as predicts the sample label using Minimum Bayesian Risk based on the Hellinger distance. This approach yielded 90.66% sensitivity, 92.52% specificity and F2-score of 0.9055 in EEG segment classification task. Moreover, the proposed approach achieved 98.56% sensitivity of seizures with the average latency 1.32s. The proposed method shows good performance on the epileptic EEG imbalanced data, and has great clinical significance for the auxiliary diagnosis of epilepsy.

Key words: epilepsy; time-domain feature; random projection; rotation forest; cost sensitive; minimum bayesian risk

0　引言

癫痫是一种由脑部神经元异常放电引起的慢性疾病，癫痫的发作往往会给病人带来极大的痛苦。世界卫生组织（WHO）在其 2017 年的报告中指出[1]，全世界大概有 50 万癫痫患者，其中的 25%为儿童。医生诊断癫痫最直接的方式为人工判读脑电图，这一过程是非常耗时的，并且给医生带来了较大的工作负担。因此，提出一种高效的、精准的癫痫脑电自动检测的算法对于癫痫的辅助诊断有着极大的临床意义。

癫痫患者的脑电分为发作期与发作间期两部分，其特征波包括棘波、尖波、棘慢复合波与尖慢复合波。癫痫脑电检测这一课题已经引起了大量科研工作者的关注，随着近些年来机器学习方法的快速发展，癫痫脑电自动检测算法的效果也有了显著的提升。2010 年，Shoeb 等人[2]提取了脑电的时空域与功率谱特征，讨论了机器学习算法在癫痫发作自动检测中的应用效果，得到了 96%的检出率以及 4.6s 的发作检测延迟。2012 年，Martis 等人[3]提出了一种基于经验模态分解（EMD）提取脑电特征的方法，并且使用 C4.5 决策树作为分类模型，得到了 95.33%的准确率。2014 年，Chen 等人[4]提出了基于小波分解的极限学习机癫痫发作检测模型，比较了样本熵，近似熵与递归定量分析（RQA）等非线性特征的分类效果。2015 年，Donos 等人[5]提取了多个维度的信号特征，并且将随机森林应用于癫痫分类中，得到了 93.84%的灵敏性，3.03s 的检测延迟。随着深度学习的兴起，2017 年，Acharya 等人[6]将脑电信号分为发作期，发作前期与正常脑电信号三部分，提出了深度卷积神经网络的癫痫发作的多分类算法。在癫痫预测领域，Khan 等人[8]

收稿日期：2018-07-25；**修回日期：**2018-09-10　**基金项目：**国家自然科学基金资助项目（61201124）；中央高校基本业务资金资助项目（222201817006）

作者简介：卫作臣(1991-)，男，博士研究生，主要研究方向为生物电信号处理、机器学习；邹俊忠(1960-)，男（通信作者），教授，博士，主要研究方向为脑电信号处理与康复、模式识别与人工智能（jzhzou2015@sina.com）；张见(1976-)，男，工程师，博士，主要研究方向为生物电信号处理、模式识别；陈兰岚（1983-），女，副教授，博士，主要研究方向为脑电信号处理、生物电信号系统控制.

基于小波分量与卷积神经网络得到了 87.8%的癫痫发作检出率以及 0.142/小时的误判率。

在以往的大量有关癫痫脑电自动检测的研究中，时域、频域、时频域以及非线性特征均有所涉及。其中时域特征是最本质的特征，医生在临床上判读脑电图中的癫痫特征波的依据往往是时域波形。时域特征相比其他特征有着以下三点优势：a）时域特征有着明确的物理意义；时域特征是最直接的特征，是无须作变换就可以观测到的信息；无须作信号平稳的假设。在之前的研究中[8~10]提出了一种基于视觉组织原则的信号波形处理方法，这种方法通过学习医生判读脑电图的视觉感知与完形的过程来建立模型，有效而灵活地分割了信号的周期，突出了时域波形状特征并去除了冗余信息，这一方法已经在癫痫放电检测、疑似癫痫的噪声自动分离以及睡眠自动分期中体现了良好的性能。

在癫痫脑电自动检测这一问题中，可采集到的癫痫波的数量是大大少于正常脑电样本的，因此实际上这是一个不平衡二分类的问题，而不平衡数据集训练出的模型一般会偏向多数类，此时的模型准确率无法很好地评价模型性能[11]。在数据层面上，不平衡数据集可以使用欠抽样或过抽样方法来使其变得平衡，但这两种方法都有着各自的缺陷。欠抽样方法会减少模型学习到的多数类样本信息，而过采样方法通过随机复制少数类样本来实现数据集的平衡，这样会导致模型的泛化性能变差，容易过拟合。SMOTE[12]算法通过近邻来生成不重复的少数类样本，但容易造成类边界模糊。代价敏感（cost-sensitive）思想是另一种解决不平衡数据集问题的方法，它是针对不同类错分代价不同所提出的模型学习思想，这一学习方法不关注模型的错误率，而是通过预测的最小代价来训练与选择模型。Domingos[13]提出了基于元代价（Metacost）的代价敏感算法，即将基分类器当作黑箱子，通过代价敏感矩阵得出的误分类风险来更新训练集标签。在分类问题上，大多数机器学习方法都假设各类错分代价相同，而对于不平衡数据集则不是这样，少数类样本比多数类样本珍贵，并且应当对少数类的错分施加更大的惩罚，因此代价敏感思想是一种在算法层面上处理不平衡数据集的方法。

本文的提出了一种快速、高效的不平衡分类算法，并脑电信号中癫痫波的自动检测问题进行分析，所使用的数据集为不平衡数据集，以此来研究在类别不平衡情况下所提出的算法自动识别癫痫脑电的性能。本文提出了基于视觉组织原则的时域信号特征提取方法，算法部分主要的贡献与创新点可分为以下两部分：

a)改进了旋转森林算法[14]，使用随机映射代替标准旋转森林特征旋转过程中的 PCA 映射来提高运算的时间效率，并且 Johnson-Lindenstrauss Lemma 定理保证了随机降维方法的精度。

b)提出了基于贝叶斯最小风险的代价敏感分类方法，根据海林格距离来评估样本不平衡程度，并赋予少数类和多数类不同的错分代价。根据代价矩阵计算出样本预测风险，从而将预测样本分为误分类风险最小的那一类。相比传统方法，这一基于贝叶斯最小风险的方法有效地提高了少数类即癫痫波的检出率。

1 方法与模型

1.1 时域特征提取

1.1.1 基于视觉组织原则的周期分割法

这一简单高效的时域算法在之前的研究中[8]已被详细介绍，在这一小节简单陈述算法流程与相关概念。

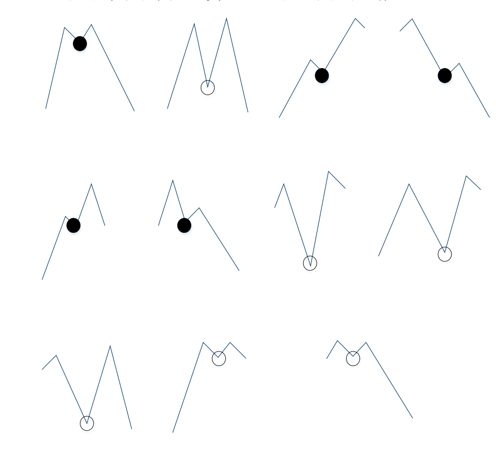


图 1 相邻波形的 11 种组合模式,空心圆表示完整的波形组合,实心圆为不完整的波形组合

定义 1 增减序列。定义 $s(i)$ 为时序列的第 i 个采样点，假设 $s(i)$， $s(i+m)$ 为两个局部极小值， $s(i+n)$ 为一个局部极大值，则认为这三个点构成了一个单位波，i 到 $i+n$ 为上升序列，i 到 $i+m$ 为下降序列。

1)杂波合并条件

首先，由于癫痫特征波中的棘波为 20~70 ms，尖波周期为 70~200 ms，所以周期小于 20 ms 的波被认为是杂波；其次，波的幅值小于 10 μV 同样被判定为杂波。

a) 不完整波形条件

定义序列 $s(i)$ 中的极值点索引为序列 $s(a(i))$ ，那么假设相邻两个波的极值序列为 [$s(a(i))$, $s(a(i+1))$, $s(a(i+2))$, $s(a(i+3))$, $s(a(i+4))$],这两个波的组合模式共 11 种，如图 1 所示，其中 5 种为不完整的模式，需要合并。令

$$
\begin{aligned}
h_1 &= s(a(i+1)) - s(a(i)) \\
h_2 &= s(a(i+1)) - s(a(i+2)) \\
h_3 &= s(a(i+3)) - s(a(i+2)) \\
h_4 &= s(a(i+3)) - s(a(i+4))
\end{aligned}
$$

则图 1 中不完整的波形模式可归纳为下述 5 种规则：

$$h_1 / h_2 < r \ \& \ h_3 / h_4 < r$$

$$h_2 / h_1 < r \ \& \ h_4 / h_3 < r$$

$$h_2 / h_1 < r \ \& \ h_3 / h_4 < r$$

$$h_2 / h_1 < r \ \& \ h_3 / h_4 \geq r \ \& \ h_4 / h_3 \geq r \ \& \ h_2 / h_3 < r$$

$$h_3/h_4<r\ \&\ h_1/h_2\ge r$$
$$\&\ h_2/h_1\ge r\ \&\ h_3/h_2<r$$

其中 r 为合并系数，之前的研究中表明 $r=0.5$ 可以较好地适应各种节律的波。对于尖形波，$r=0.3$，尖形波需要满足以下三个条件：$20ms\le T\le 200ms$；$h/T>2$；突出于背景：$h/h_{pre}>4$，其中 h_{pre} 为该波之前 1 分钟内所有波的平均幅值。

b) 增减序列合并算法（MIDS）

每两个相邻的基于极值点定义出的波做一次运算，根据规则评价是否为不完整波或杂波，若是，则执行算法 1 的合并操作。

算法 1 增减序列合并（MIDS）

Input：raw time sequence s；

index of extremum sequence a.

Output：the merged sequence.

for each value $s(a(i))$ of s DO

if [$s(a(i))$, $s(a(i))$,$s(a(i+2))$] is a cluster or incomplete then

$a(i+1)\leftarrow argmax(s(a(i+1)),s(a(i+3)))$

$a(i+1+k)\leftarrow a(i+3+k),\ k=1\ldots N-3-i$

end if

$i\leftarrow i+2$

end for

return $s(a)$

1.1.2 特征计算

使用上节的 MIDS 算法处理原始信号后，得到了周期分割序列，以分割后的单个完整波形[$s(a(i))$, $s(a(i+1))$, $s(a(i+2))$] 为基础，提取 17 维时域特征，各个特征计算过程如下：

a)周期,即 $T=a(i+2)\text{-}a(i)$；

b)上升幅度 $h_{in}=s(a(i+1))-s(a(i))$

c)下降幅度 $h_{de}=s(a(i+1))-s(a(i+2))$

d)上升时长 $a(i+1)-a(i)$；

e)下降时长 $a(i+1)-a(i+2)$；

f)波形幅值的标准差 $\sqrt{\frac{1}{T}\sum_{k=a(i)}^{a(i+2)}(s(k)-s_{mean})^2}$，评价波形变化程度;

g)锐度 $2s(a(i+1))-s(a(i))-s(a(i+2))$；

h)幅值平方和 $\sqrt{\sum_{k=a(i)}^{a(i+2)}s(k)^2}$；

i)幅值均值 $\frac{1}{T}\sum_{k=a(i)}^{a(i+2)}s(k)$；

j)相邻点差值标准差 $\sqrt{\sum_{k=a(i)}^{a(i+2)-1}(s(k+1)-s(k))^2}$，评价波形波动程度;

k)Hjorth 参数[15]，复杂度，化简表示为

$$\frac{\sqrt{\sum_{k=a(i)}^{a(i+2)-2}(s(k+1)-s(k))^4}\sqrt{\sum_{k=a(i)}^{a(i+2)}(s(k))^2}}{\sum_{k=a(i)}^{a(i+2)-1}(s(k+1)-s(k))^2}$$

l)峰度系数 $\frac{T\sum_{k=a(i)}^{a(i+2)}(s(k)-s_{mean})^4}{(\sum_{k=a(i)}^{a(i+2)}(s(k)-s_{mean})^2)^2}-3$，度量波形的聚集程度;

m)最大幅值与波形标准差的比值 $\frac{max(s(k))}{\sqrt{\frac{1}{T}\sum_{k=a(i)}^{a(i+2)}(s(k)-s_{mean})^2}}$；

n)上升角度的余切 $\frac{a(i+1)-a(i)}{s(a(i+1))-s(a(i))}$；

o)下降余切 $\frac{a(i+1)-a(i+2)}{s(a(i+1))-s(a(i+2))}$；

p)幅值较大侧/周期 $\frac{max(h_{in},h_{de})}{T}$；

q)幅值较小侧/周期 $\frac{min(h_{in},h_{de})}{T}$。

在本实验中，选择时长为 1 s 的多导联脑电片段为一个样本，因此每一个导联中的这 17 个时域特征取所有完整波的均值，因此特征维度为 17*N，N 为导联数量。

1.2 基于随机映射的旋转森林

1.2.1 随机映射

随机映射（random porjection, RP）是机器学习中的一种降维方法，它将高维空间 R^d 中的样本集 S_d 通过一个随机转换矩阵 $M_{k,d}$ 映射到低维空间 R^k 中，如式（1）[16]所示，Johnson-Lindenstrauss Lemma[17]给出了这种映射方法降维后的精度范围。

$$S_k=M_{k,d}S_d \tag{1}$$

定理 1 Johnson-Lindenstrauss Lemma。高维空间中的 n 个样本点可以映射到一个为 $O(log\ n/\varepsilon^2)$ 的低维空间中，并且样本间距离变化不会超过 $(1\pm\varepsilon)$。

对于任意的 $0<\varepsilon<1$ 与正整数 n，有整数 k

$$k\ge 4(\varepsilon^2/2-\varepsilon^3/3)^{-1}\ln n \tag{2}$$

则对于任意的空间 R^d 中包含 n 个样本的数据集 S，存在映射：$f:R^d\to R^k$，且对于任意 $u,v\in S$ 存在：

$$(1-\varepsilon)\|u-v\|^2\le\|f(u)-f(v)\|^2\le(1+\varepsilon)\|u-v\|^2 \tag{3}$$

证明过程见文献[17]。

因此，定理 1 指出，在低维空间维度不小于一个阈值的前提下，映射后的样本间距离相比原始空间中的数据集在一个范围内保持不变，并且映射 f 可以随机得出。相比其他的降维方法，随机映射的计算量非常小，对于 n 个样本点的原始数据集，随机映射的复杂度仅为 $O(dkn)$。其次，转换矩阵为随机产生，无须计算原始数据集的信息，仅通过保证样本间距离变化误差来实现降维，这一特性对于样本间距离有意义的数据集有着较大优势。

近几年来，随机映射凭借着它较小的计算量以及相比传统降维方法较好的性能已被应用在各种领域中[18,19]。本文使用随

机映射来实现旋转森林算法中的特征旋转部分，在保证变换效果的前提下，提高了整个算法的计算效率。

1.2.2 改进的旋转森林算法

旋转森林（rotation forest）是Rodrı´guez[20]于2006年提出的一种基于随机森林（random forest）思想的改进算法。旋转森林在抽取样本子集的同时，将特征分为多个子集，每一个基分类器都选取不同的特征子集进行特征维度线性变换，以此增加基分类器间的多样性。相比传统的随机森林算法，旋转森林由于其基分类器更大的差异，相比随机森林会进一步地降低各个子模型间的相关性，从而得到一个方差更低的模型，这一算法在部分领域已经得到应用[21]。本文使用随机映射代替文献[20]中的主成分分析（PCA）来优化旋转森林算法中的核心部分，即特征线性变换的过程。随机映射相比PCA不需要计算数据的协方差矩阵，并且定理1保证了随机投影的精度，整个计算流程如下：

a)有训练集 X_{n*N}，其中 n 为特征维度，N 为样本数量，Y 为标签列，E_k 为第 k 个基分类器。假设特征分为 m 个不相交特征子集，则每个子集包含特征 $F=n/m$ 个；

b)随机抽取样本训练基分类器，使用随机映射进行特征的线性变换，按特征索引顺序排列后得到每个基分类器的线性变换矩阵 R_k；

c)使用数据集 XR_k,Y 训练基分类器 E_k；

d) 则预测时样本各个类别的输出概率为 $p(j|x)=\frac{1}{K}\sum_1^K pred_k(j|xR_k)$，其中 j 为类别，样本最终预测类别为 $\arg\max(p(j|x))$。

1.3　贝叶斯最小风险预测

1.3.1 海林格距离

海林格距离（Hellinger distance）由Hellinger提出，在统计学中，用来度量两个概率分布的相似程度，是 *f*-divergence 散度的一种，定义如下：

定义2　海林格距离（HD）。有概率分布P和Q，则这两个分布之间的海林格距离为

$$h(P,Q)=\frac{1}{\sqrt{2}}\left\|\sqrt{P}-\sqrt{Q}\right\|_2 \tag{3}$$

Cieslak等人基于海林格距离提出了海林格决策树（HDDT）算法[22]，通过在决策树分裂时计算节点上样本类间的海林格距离代替传统的gini指数，在不同样本比例的数据集上取得了良好的效果。由此可知，相比直接使用样本比例度量类间分布情况的方法，海林格距离度量表现出了更好的性能。

1.3.2 贝叶斯风险计算

本实验中所使用的为不平衡癫痫脑电数据集，随着类间比例的变化，两类样本各自的误分类代价应当随之而变化。本文使用代价敏感思想[13]来研究这一问题，提出了一种基于海林格距离的新颖的类间不平衡评价方式，并根据评价值给出代价敏感损失矩阵，进而计算出各类别贝叶斯最小风险实现预测。

定义3　HD评价指标。有样本集 S，包含两类标签，则类间的海林格距离为

$$h(+,-)=\frac{1}{\sqrt{2}}\left\|\sqrt{\frac{S_+}{S}}-\sqrt{\frac{S_-}{S}}\right\|_2 \tag{4}$$

其中：S_+表示正样本数量，S_-表示负样本数量，S 表示样本集总的样本数量。

定义4　代价矩阵。对于样本集 S，二分类问题的误分类代价矩阵为

$$C=\begin{bmatrix}0 & C_{-,+}\\ C_{+,-} & 0\end{bmatrix} \tag{5}$$

其中：$C(-,+)$，表示将负类预测为正类的代价，反之亦然，此处的 C 值根据类间的海林格距离算出。

定义5　贝叶斯最小风险预测。令 $p(j|x)$ 为模型将样本 x 预测为 j 类概率值，则最终预测类别为

$$y=\arg\min_i(\sum_j p(j|x)C(j,i)) \tag{6}$$

本文提出的基于贝叶斯最小风险预测的算法考虑了训练模型的数据集的样本分布情况，根据类间的不平衡程度给出代价损失矩阵，将使用不平衡数据集训练出的偏向多数类的模型通过赋予不同的误分类损失加以调节。这一算法将模型看做一个黑箱子，不在模型内部针对不平衡训练集问题进行调整，而是在模型之外根据训练集的实际分布给出预测概率的置信度，通过计算最小贝叶斯风险来预测样本类别，从而减少不平衡训练集带来的影响。

算法2：基于海林格距离的贝叶斯最小风险预测（HD-MBR）

Input：Train set S；

Probabilities of each class for sample x $p(j|x)$.

Output：predicted class y.

calculate Hellinger Distance:

$$HD\leftarrow\frac{1}{\sqrt{2}}\left\|\sqrt{\frac{S_+}{S}}-\sqrt{\frac{S_-}{S}}\right\|_2$$

calculate cost matrix:

$$C\leftarrow\begin{bmatrix}0 & 1-\sqrt{2}HD\\ 1 & 0\end{bmatrix}$$

The predicted class:

$$y=\arg\min_i(\sum_j p(j|x)C(i,j))$$

return y

1.4　算法流程

算法流程如图2所示。

2　实验结果与讨论

2.1　数据集

本文使用公开的癫痫脑电数据集CHB-MIT[23~26]来测试所提出算法的性能。为了确保数据集特征维度的统一，选用数据集中使用23导联采集脑电信号的12个病人的数据，23个导联

包括 FP1-F7，F7-T7，T7-P7，P7-O1，FP1-F3，F3-C3，C3-P3，P3-O1，FP2-F4，F4-C4，C4-P4，P4-O2，FP2-F8，F8-T8，T8-P8，P8-O2，FZ-CZ，CZ-PZ，P7-T7，T7-FT9，FT9-FT10，FT10-T8，T8-P8。信号采样率为 256Hz，每一份病人的数据中均包含多次癫痫发作。

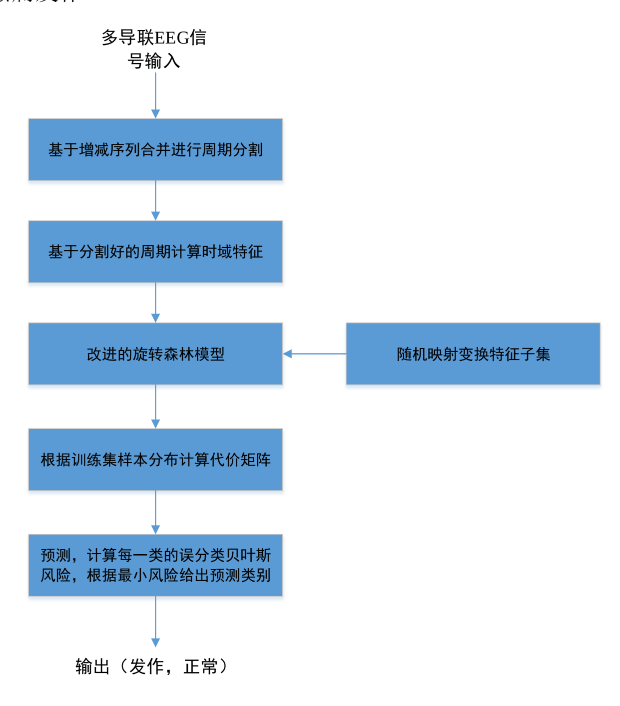


图 2　算法整体流程

2.2　实验设计

实验环境为 Ubuntu1 6.04 系统、英特尔 i7-7700k 处理器、32 GB 运行内存。实验数据中样本为时长 1 s 的多导联脑电片段，取所有的癫痫发作脑电片段构造正样本，并且在每一次癫痫发作之前的脑电信号中随机选出时长固定的间期脑电片段。由于发作时长不固定，因此每次发作的正负样本比例是不同的。按此方法处理每一次发作，得到一个不平衡的数据集，不同份数据的不平衡比例约在 1：4 到 1：20 之间。为了更好地评估所提出算法的性能，使用留一法（leave-one-out）进行交叉测试。即假定患者 n 的脑电数据中存在 N 次发作，则测试第 i 次发作时使用剩下的 N-1 次数据训练，迭代 i 的值直至 i=N。

算法性能评价分为两部分，分别为

a)基于片段评价（segment-based）。即以 1s 时长的样本为单位，评估模型性能，是一个癫痫脑电样本的分类问题。

b)基于事件评价（event-based）。以一次发作为单位，判断是否检出以及检出延迟，是一个癫痫发作的检测问题。

由于样本集正负比例不平衡，因此无法用准确率评价模型性能，所选取的评价指标为灵敏性（sensitivity），特异性（specificity），$F\beta$ 分数（$F\beta$ -score）以及检测延迟（latency），指标计算方式如下：

$$sensitivity = TP / (TP + FN)$$
$$specificity = TN / (TN + FP)$$
$$precision = TP / (TP + FP)$$
$$F\beta = (1+\beta^2)\frac{precision * sensitivity}{(\beta^2 * precision) + sensitivity}$$
$$latency = seizure\ onset - predicted\ seizure\ onset$$

其中：检测延迟为基于事件的评价方式的评价指标。$F\beta$ 分数是精准率与灵敏性的一个加权平均，由于在癫痫检测中，漏检一次癫痫比误判一次正常脑电的后果要严重，因此实验中选择 $\beta = 2$ 来表示癫痫样本的重要性，即使用 $F2$ 分数作为评价指标。

2.3　结果分析

实验对 12 份数据各自的癫痫发作做交叉测试，基于 1s 癫痫样本的分类结果如表 1 所示。

表 1　基于 1 s 片段的癫痫脑电分类结果

编号	灵敏性/%	特异性/%	F2 分数
1	94.59	97.36	0.9501
2	82.98	59.13	0.7655
3	92.52	97.55	0.9623
4	88.60	94.66	0.8959
5	92.75	98.21	0.9377
6	96.72	90.04	0.9117
7	90.80	96.00	0.9171
8	93.12	94.58	0.9331
9	97.03	93.27	0.9601
10	87.43	98.24	0.8912
11	97.02	97.11	0.9700
12	74.33	94.05	0.7718
总体	90.66	92.52	0.9055

从表 1 中可以看出，模型在所有数据上均表现出了良好的效果，大部分灵敏性在 85%以上，特异性在 90%以上。平均灵敏性和特异性 90.66%、92.52%，平均 F2 分数为 0.9055。

为了体现所提出算法相比传统机器学习算法的优越性能，使用随机森林以同样的方式做测试，所提出算法正是在随机森林上进行了特征子空间分割、特征旋转、代价敏感以及贝叶斯最小风险预测这些改进得到的。在评价指标中，F2 分数综合考虑了模型的灵敏性和精准率，图 3 通过不同病人数据的 F2 分数来进行算法之间的性能比较。同时，表 2 列出了随机森林在 12 份数据上的平均指标值，结果表明，提出的算法相比随机森林，提高了 10.29%的灵敏性，0.0867 的 F2 分数，而特异性只下降了 2.8%。在临床上，漏判的癫痫放电相比误判的正常脑电要付出更大的代价，提出的算法则符合这个思想，并且相比传统模型提升明显。

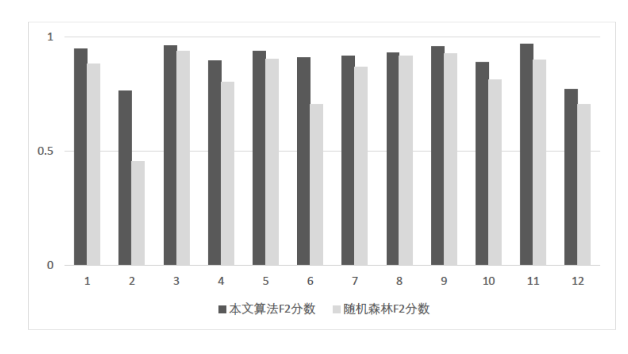


图 3 模型 F2 分数比较

表 2 与传统模型的结果对比

模型	灵敏性/%	特异性/%	F2 分数
随机森林	80.37	95.32	0.8188
本文算法	90.66	92.52	0.9055

进一步地，为了验证所提出算法中的核心部分贝叶斯最小风险预测相比传统不平衡分类算法的优越性能，在同样的模型下，分别采用权重法与 SMOTE 过采样这两种方法处理训练集，来代替本文的贝叶斯最小风险预测方法。在 SMOTE 处理过程中，调节采样率生成与多数类等量的少数类样本，在权重法中，则按照样本比例加权。从表 3 的对比结果中可以看出，本文算法相比 SMOTE 方法与权重法在灵敏度上分别提高了 3.85%、4.98%，F2 分数提高了 0.0285、0.0384。这一结果表明，本文提出的算法在不平衡数据集上的效果要优于传统的权重法与 SMOTE 算法。

表 3 与传统不平衡数据集处理算法的结果对比

算法	灵敏性/%	特异性/%	F2 分数
权重法	85.68	93.94	0.8671
SMOTE	86.81	93.63	0.8770
本文算法	90.66	92.52	0.9055

为了进一步评估模型的发作检测性能，将每次测试的发作脑电按时间顺序排列，以此来观察每次发作是否被检出，以及其检出延迟，表 4 列出了基于事件的癫痫检测结果。

表 4 癫痫发作检测结果

编号	检出率/%	延迟/s
1	100	1.14
2	100	0.5
3	100	0.86
4	100	2.25
5	100	2.2
6	100	0.5
7	100	0.67
8	100	1.6
9	100	1.25
10	100	1.57
11	100	0.57
12	92.31	2.77

检测结果表明，11 份数据中的发作全部被检出，仅在数据 12 中存在漏检。考虑每一份数据的发作次数，平均检出率为 98.65%，平均检测延迟为 1.32s。

由于现阶段提出的大部分算法的评价方式为基于癫痫事件的评价，表 5 列出了在公开数据集 CHB-MIT 上的其他癫痫发作事件检测研究的结果，并与本文所提出方法的比较。结果表明，所提出的算法效果相比其他文献有所提高，并且与大部分其他的研究不同，本文的实验结果是在不平衡数据集上得到的，这说明该算法在不平衡数据集上有着极大的潜力。

表 5 与其他方法比较结果

方法	检出率/%	延迟/s
文献[2]方法	96	4.6
文献[25]方法	100.0	3.36
文献[26]方法	98.5	1.76
本文算法	98.65	1.32

2.4 讨论

大部分脑电信号的模式识别研究均是在平衡数据集上测试所提出的算法性能，本实验考虑患者脑电信号中的癫痫波数量很少，针对这一不平衡的数据集分类问题提出了基于贝叶斯风险预测的算法，并在不平衡的癫痫脑电数据集上进行应用分析，由结果可以看出，在不同类别比例的测试集上该算法均可以保持良好的性能以及稳定的效果。只要知道训练该模型的训练集样本分布情况，就可以通过不同的误分类代价计算贝叶斯风险，以此代替传统的输出概率实现样本的分类。

本文使用了高效简洁的时域特征，在医生判读脑电图的过程中，信号的时域特征是最直接的特征，且物理意义明确，医生往往可以通过癫痫波中的尖、棘波及其复合波的形状特征做出判断。本文使用了我们之前研究提出的周期分割算法来进行时域特征的提取，这一算法基于合并、分割的规则模拟医生在判读时域信号的过程中视觉完形，神经感受野的概念，在时序列信号处理中相当于为特征提取的过程给出了一个灵活、自适应时间窗口，该窗口符合人的视觉感知与判断，所分割出的完整波形符合人的视觉感知，因此提取出的特征也提供了时序信号中每一个完整波形的信息，实验结果证明了所提取的特征的有效性。另外，本文选用了 1 s 时长的片段作为样本进行分类，相比大部分癫痫脑电识别算法，这是一个非常短的时间窗口，大部分算法不足以在 1 s 内提取出判别癫痫放电的信息，而对于基于周期分割的特征提取方法来说，这些信息是足够的。进一步来看，这一特征提取方法是基于每一个波来提取特征的，从而有效地判断出一个完整的波形是否是异常放电，这种在短时间内提取有效特征的方法在在线检测中有着极大的优势。

在模型部分，本文使用随机映射优化了传统旋转森林方法中特征子空间的变换这一核心部分。相比传统的 PCA 旋转方法，随机映射是一种快速的投影算法。PCA 需要计算特征的协方差矩阵，并通过 SVD 分解来求得该矩阵的特征值与特征向量，排序后得出变换矩阵，这一计算计算代价高昂。随机映射

则使用随机生成的变换矩阵进行映射，且定理 1 给出的距离不变性可以保证变换后的精度，相当于根据数据间的距离进行投影。随机映射无需计算数据的协方差矩阵及其特征值、特征向量，计算量要小于 PCA。

贝叶斯最小风险的预测方法为不平衡数据集训练出来的模型提供了良好的分类性能，并且计算过程简单。表 2 的结果表明，对比未作处理的传统随机森林算法在效果上有着明显的提升。传统的不平衡分类方法往往为过采样、欠采样等，本文将所提出的不平衡分类算法与权重法、SMOTE 算法对比，表 3 已经从效果上体现了本文算法要优于这两种传统算法，同时在训练效率上，本文算法在时间和空间效率上均有所提升。在本文的树模型中，权重法对少数类错分做出的加权惩罚体现在每一个叶节点分裂的过程中，所需的乘法次数要多于本文算法。SMOTE 算法根据所有的少数类样本及其近邻来生成少数类样本，需要根据欧式距离计算每一个少数类样本的近邻，并根据预设好的采样率与每一个近邻生成新的样本。假设数量为 n 的训练集中，少数类样本数量为 p，SMOTE 算法在计算每个少数类样本的 K 近邻时需要遍历 p^2 次，存储近邻索引则需要 p*k 的空间。在生成样本过程中，时间复杂度则与需要生成的样本数目线性相关，并且需要额外的空间去存储这些生成的样本。本文算法的在知道各类样本数量的情况下，训练过程与平衡数据集情况下没有区别，除计算一次海林格距离外不需要进行其他的额外计算，仅需在预测时遍历一次测试集即可得出最小风险，这一过程的时间效率高于 SMOTE 算法。所需额外存储的变量仅为代价敏感矩阵，而在二分类问题中，代价敏感矩阵为两个值，所需存储空间可忽略不计，远低于 SMOTE 方法的空间复杂度。

从理论角度来看，本文算法不需要对原始数据集做生成、更新或者欠采样处理。欠采样方法会丢弃掉大量的负样本信息，影响模型性能；随机过采样采用随机复制少数类样本来增加少数类样本，会造成样本多样性不足，少数类样本的信息实际上没有增加，造成模型泛化性能变差；权重法对少数类样本加权的过程相当于产生了新的数据分布，将分类器的重点集中在少数类样本的身上，但这个权值的选择往往过于主观；SMOTE 过采样的方法是通过近邻合成新样本，为生成的样本提供了一定的多样性，但同样存在缺陷，首先最优近邻数值难以确定，其次由类边缘样本生成的样本会造成边界模糊的问题，加大样本集的可分性。本文提出的不平衡分类算法统计了数据的分布情况，但不需要对数据集做任何修改，而是根据分布给出不同的误分类代价，在模型层面对不同类别的错分情况计算风险，通过获得一个整体误分类风险最小的模型来实现不平衡分类问题的解决。

本文的算法框架由基于视觉组织的时域特征提取、改进的旋转森林以及贝叶斯最小风险预测三部分组合而成。所提取的时域特征为短时间的波形特征，如时长、幅值、锐度、峭度等等，这些特征大部分都是可以视觉观察到的一些波的形状特征，从医生视觉判读的角度来看，这些特征已经为单个波提供了一定的可分性。模型部分采用随机映射优化了标准的旋转森林算法，并且针对特征的线性变换降低了各个子分类器的相关性，可以进一步地降低整个模型的方差。基于贝叶斯最小风险的预测则是本文算法中解决数据不平衡的核心技术，效果上优于传统模型以及常见的不平衡分类算法，并且训练效率上有所提升。因此整个算法框架在不平衡数据集的分类问题上表现出了优越的性能。

3　结束语

本文提出了基于时域特征的不平衡癫痫脑电检测算法，将多导联脑电信号分割周期后的时域特征输入改进的旋转森林模型，并且使用贝叶斯最小风险代替输出概率得到测试样本的类别。最终在基于 1s 的癫痫放电分类问题上得到了 90.66%的灵敏性，92.52%的特异性，0.9055 的 F2 分数，相比传统的随机森林模型，提高了 6%的灵敏性，0.9 的 F2 分数，并且在发作检测方面，检出了 98.56%的癫痫发作，检测延迟为 1.32s。这一算法在不平衡脑电数据集上得到了优良的性能，优于许多平衡数据集上的自动检测算法，并且解除了癫痫样本不充足这一限制。本实验中选取固定时长的间期脑电片段构建不同类别比例的样本集，若加大不平衡比例，本文提出的方法相比传统方法理论上提升会更显著。同时算法符合临床判读脑电图的思想，有着极大的临床意义。

参考文献：

[1] World Health Organization. Epilepsy [R] , Geneva: WHO, 2017.

[2] Shoeb A, Guttag J, Application of machine learning to epileptic seizure onset detection [C]// Proc of the 27th, International Conference on Machine Learning. 2010.

[3] Martis R J, Acharya U R, Tan J H, *et al.* Application of empirical mode decomposition (EMD) for automated detection of epilepsy using EEG signals [J]. International Journal of Neural Systems, 2012, 22 (6): 1250027.

[4] Chen Lanlan, Zhang Jian, Zou Junzhong, *et al.* A framework on wavelet-based nonlinear features and extreme learning machine for epileptic seizure detection [J]. Biomedical Signal Processing & Control, 2014, 10 (1): 1-10.

[5] Donos C, Dümpelmann M, Schulzebonhage A. Early seizure detection algorithm based on intracranial EEG and random forest classification [J]. International Journal of Neural Systems, 2015, 25 (05): 150426195043004.

[6] Acharya U R, Oh S L, Hagiwara Y, *et al.* Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals [J]. Computers in Biology & Medicine, 2017.

[7] Khan H, Marcuse L, Fields M, *et al.* Focal onset seizure prediction using convolutional networks [J]. IEEE Trans on Biomedical Engineering, 2018, PP (99): 1-9.

[8] Zhang Jian, Zou Junzhong, Wang Min, *et al.* Automatic detection of

interictal epileptiform discharges based on time-series sequence merging method [J]. Neurocomputing, 2013, 110 (8): 35-43.

[9] Wei Zuochen, Zou Junzhong, Zhang Jian. Automatic recognition of chewing noises in epileptic EEG based on period segmentation [J]. Neurocomputing, 2016, 190: 107-116.

[10] 李同庆，邹俊忠，张见，等. 基于周期分割的睡眠自动分期研究 [J]. 计算机工程与应用, 2018. (Li Tongqing, Zhou Junzhong, Zhang Jian, *et al.* The Research of automatic staging of sleep based on period segmentation [J]. Computer Engineering and Applications, 2018, .)

[11] Zhang Chong, Tan K C, Li Haizhou, *et al.* A cost-sensitive deep belief network for imbalanced classification [J]. IEEE Trans on Neural Networks & Learning Systems, 2018, PP (99): 1-14.

[12] Ma Li, Fan Suohai. CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests [J]. Bmc Bioinformatics, 2017, 18 (1): 169.

[13] Domingos P, MetaCost: a general method for making classifiers cost-sensitive [C]// Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 1999: 155-164.

[14] Su Chong, Ju Shenggen, Liu Yiguang, *et al.* Improving random forest and rotation forest for highly imbalanced datasets [J]. Intelligent Data Analysis, 2015, 19 (6): 1409-1432.

[15] Olejarczyk E, Jozwik A, Zmyslowski W, *et al.* Automatic detection and analysis of the EEG sharp wave-slow wave patterns evoked by fluorinated inhalation anesthetics [J]. Clinical Neurophysiology Official Journal of the International Federation of Clinical Neurophysiology, 2012, 123 (8): 1512-1522.

[16] Bingham E, Mannila H. Random projection in dimensionality reduction: applications to image and text data [C]// Proc of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2001: 245-250.

[17] Dasgupta S, Gupta A. An elementary proof of the Johnson-Lindenstrauss Lemma [J]. Random Structures & Algorithms, 1999, 22 (1) .

[18] Li Gen, Gu Yuantao. Restricted isometry property of Gaussian random projection for finite set of subspaces [J]. IEEE Trans on Signal Processing, 2018, PP (66): 1705-1720.

[19] Hoyos-Idrobo A, Varoquaux G, Thirion B. Fast brain decoding with random sampling and random projections [C]// Proc of International Workshop on Pattern Recognition in Neuroimaging. 2016: 1-4.

[20] Rodríguez J J, Kuncheva L I, Alonso C J. Rotation forest: A new classifier ensemble method [J]. IEEE Trans on Pattern Anal Mach Intell, 2006, 28 (10): 1619-1630.

[21] 陆慧娟，刘亚卿，孟亚琼，等. 面向基因数据分类的核主成分分析旋转森林算法 [J]. 计算机科学与探索，2017, 11 (10): 1570-1578. (Lu Huijuan, Liu Yaqing, Meng Yaqiong, *et al.* Classifier algorithm of genetic data based on kernel principal component analysis and rotation forest [J]. Journal of Frontiers of Computer Science and Technology, 2017, 11 (10): 1570-1578.)

[22] Cieslak D A, Chawla N V. Learning decision trees for unbalanced data [C]// Proc of European Conference on Machine Learning and Knowledge Discovery in Databases. Springer-Verlag, 2008: 241-256.

[23] Shoeb A. Application of machine learning to epileptic seizure onset detection and treatment [D]. Cambridge: Massachusetts Institute of Technology, 2009.

[24] 王凤琴，卢官明，柯亨进，等. 基于跨层全连接神经网络的癫痫发作期识别 [J/OL]. 计算机应用研究, 2019, 36 (7) . [2018-04-12]. http://www.arocmag. com/article/02-2019-07-013. html. (Wang Fengqin, Lu Guanming, Ke Hengjin. Epileptic EEG identification with cross layer fully connected neural network [J/OL]. Application Research of Computers, 2019, 36 (7) .)

[25] Supratak A, Li Ling, Guo Yike. Feature extraction with stacked autoencoders for epileptic seizure detection [C]// Proc of the 36th Annual International Conference of IEEE Engineering in Medicine and Biology Society. Piscataway, NJ: IEEE Press, 2014: 4184-4187.

[26] Ahammad N, Fathima T, Joseph P. Detection of epileptic seizure event and onset using EEG [J]. Biomed Research International, 2014, 2014 (1): 450573.